OXFORD

Briefings in Bioinformatics, 2022, **23(5)**, 1–14
https://doi.org/10.1093/bib/bbac160
Problem Solving Protocol

Biological activities of drug inactive ingredients

Chenyang Zhang†, Minjie Mou†, Ying Zhou†, Wei Zhang, Xichen Lian, Shuiyang Shi, Mingkun Lu, Huaicheng Sun, Fengcheng Li, Yunxia Wang, Zhenyu Zeng, Zhaorong Li, Bing Zhang, Yunqing Qiu, Feng Zhu and Jianqing Gao

Corresponding authors. Jianqing Gao, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China. E-mail: gaojianqing@zju.edu.cn; Feng Zhu, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China. E-mail: zhufeng@zju.edu.cn
†Chenyang Zhang, Minjie Mou and Ying Zhou contributed equally to this work.

Abstract

In a drug formulation (DFM), the major components by mass are not Active Pharmaceutical Ingredient (API) but rather Drug Inactive Ingredients (DIGs). DIGs can reach much higher concentrations than that achieved by API, which raises great concerns about their clinical toxicities. Therefore, the biological activities of DIG on physiologically relevant target are widely demanded by both clinical investigation and pharmaceutical industry. However, such activity data are not available in any existing pharmaceutical knowledge base, and their potentials in predicting the DIG-target interaction have not been evaluated yet. In this study, the comprehensive assessment and analysis on the biological activities of DIGs were therefore conducted. First, the largest number of DIGs and DFMs were systematically curated and confirmed based on all drugs approved by US Food and Drug Administration. Second, comprehensive activities for both DIGs and DFMs were provided for the first time to pharmaceutical community. Third, the biological targets of each DIG and formulation were fully referenced to available databases that described their pharmaceutical/biological characteristics. Finally, a variety of popular artificial intelligence techniques were used to assess the predictive potential of DIGs' activity data, which was the first evaluation on the possibility to predict DIG's activity. As the activities of DIGs are critical for current pharmaceutical studies, this work is expected to have significant implications for the future practice of drug discovery and precision medicine.

Keywords: drug inactive ingredient, biological activity, drug formulation, database, artificial intelligence

Introduction

In a drug formulation (DFM), the major components by mass are not Active Pharmaceutical Ingredient (API) but rather Drug Inactive Ingredients (DIGs which are also known as excipients) [1]. DIGs can usually reach much higher concentration (up to 100 times in gastrointestinal tract) than that achieved by API [2], which raises great concerns regarding their clinical toxicity [3], DIG-drug interaction [4], drug resistance [5–8], etc. For instance, due to the great concerns about the safety of DIG, any new medicine authorized in Europe has to stipulate the quantitative details of DIGs in 'Summary of Product Characteristics' [3]. Since biological activity of DIG on physiologically relevant targets is widely demanded by both clinical investigations and pharmaceutical industry [9], there is an explosive growth of the studies exploring such valuable activity information [9–16]. Particularly, various DIGs were found to interact with therapeutic targets, drug transporters, or drug metabolizing enzymes [10–13], and therefore induce variation in drug bioavailability or toxicity [14–16]. Moreover, based on *in-silico* predictions and experimental validations, a recent study systematically reveals that the molecular interaction pattern of DIG is complicated, which requires the extensively explicit descriptions on the activity information of DIGs [9].

To facilitate the pharmaceutic studies based on the DIG information [4, 17, 18], several valuable open-access databases are available, which focus on providing the maximum daily exposure to DIG [Food and Drug Administration (FDA) IID [19] and STEP [20]], DIG composition within a DFM (Drugs@FDA [21] and Pillbox [22]) and molecular structure of DIGs (Excipients Browser [23] and Excipient Raman DB [24]). Although these databases are popular for current pharmaceutic study, they do not provide any activity data for each DIG. Some other databases contain the biological activity for different molecules (such as ChEMBL [25], PubChem [26] and BindingDB [27]). However, they provide very limited amount (<100) of DIGs, which is far less than the total number (~800)

Chenyang Zhang, Minjie Mou, Ying Zhou, Wei Zhang, Xichen Lian, Shuiyang Shi, Mingkun Lu, Huaicheng Sun and Fengcheng Li are the PhD/MD candidates of the College of Pharmaceutical Sciences in Zhejiang University, China. They are interested in bioinformatics and molecular biology.
Zhenyu Zeng, Zhaorong Li and Bing Zhang are senior experts on big data/AI product and solution in Alibaba Cloud. Their current focus includes Digital Health, Health Economics and Knowledge Graph.
Yunqing Qiu is a professor at the First Affiliated Hospital in Zhejiang University, China. He specializes in precision medicine, diagnosis and treatment of liver disease and system biology.
Feng Zhu is a tenured professor at the College of Pharmaceutical Sciences in Zhejiang University, China. His research group (https://idrblab.org/) has been working in the fields of OMIC-based drug discovery.
Jianqing Gao is a professor at the College of Pharmaceutical Sciences in Zhejiang University and Westlake Laboratory of Life Sciences and Biomedicine, China. His research interests include the pharmaceutics, tissue regeneration, stem cells and nanomedicine.
Received: February 28, 2022. **Revised:** April 1, 2022. **Accepted:** April 9, 2022
© The Author(s) 2022. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bib/article/23/5/bbac160/6582006 by National Science & Technology Library user on 07 October 2022

of DIGs in all FDA-approved DFMs [28]. In other words, the vast majority (>80%) of DIG's activity data remain widely dispersed in literature [9–16]. Furthermore, for those DIGs (<100) with activity information in available databases, none of them is specified as DIG in the corresponding databases, and no relationship to any drug/disease is indicated [25–27]. Based on those available data, it is difficult to explore the DIG-dependent drug safety/efficacy and attract research interest from clinical investigation or pharmaceutical industry [9, 29–36].

Here, a new database entitled 'ACDINA: biological activities of drug inactive ingredients' was therefore developed (https://idrblab.org/acdina/). First, the comprehensive literature review on all (>2000) drugs approved by US FDA was conducted, which identified a total of 23 949 DFMs that consisted of 711 unique DIGs. These DIGs covered the very wide (>50) DIG-functional classes as defined by the US Pharmacopeia [37]. Second, the activities of these DIGs and their corresponding biological targets were systematically collected from literatures, which resulted in 351 DIGs with explicit activity against 362 physiologically relevant targets from 83 target families (ABC transporter, major facilitator, GPCR rhodopsin, etc.). Third, the mapping of DIG activities to their corresponding drugs further resulted in 23 949 DFMs with ≥1 biologically active DIGs. Fourth, all DIGs and biological targets were fully cross-linked to well-established databases (including UniProt [38], PubChem [26], TTD [39], NCBI Gene [26], VARIDT [40], ChEMBL [25], Cellosaurus [41], INTEDE [42], etc.) to facilitate the prediction of drug safety/sensitivity and the assessment of DIG–drug interaction. Finally, various machine learning (ML) techniques were applied to assess the predictive potential of those newly collected data, which gave the first assessment on the possibility to predict DIG's biological activity.

All in all, the ACDINA is unique in (i) describing the largest number of well-defined DIGs and DFMs that are manually curated and carefully confirmed using their corresponding drugs, (ii) providing, for the first time, the comprehensive biological activity data for both DIGs and DFMs and (iii) fully referencing the biological targets by cross-linking them to available databases that describe their pharmaceutical/biological characteristics. As the activity data of DIGs are critical for current pharmaceutical research and industry, ACDINA is expected to have great implications for the future practice of drug discovery and precision medicine [43, 44]. The ACDINA can be freely accessed by all users at: https://idrblab.org/acdina/.

Materials and Methods

Database implementation and access of DIG-related information

ACDINA is deployed on a web server running Ubuntu v16.04.3 LTS operating system, Apache HTTP web-server 2.2.15 and Apache Tomcat servlet container. Its web interface was developed based on PHP 7.1 and Drupal 8.3.9. A variety of Drupal modules were utilized in both data call and data presentation process. In ACDINA, several dozens of data tables were stored in MySQL v15.1 to facilitate the customized database search. All ACDINA data are searchable and can be readily accessed and retrieved using diverse browsers such as Google Chrome, Mozilla Firefox, Safari and Internet Explorer 10/later. To make the access/analyses of ACDINA data convenient for users, the collected raw data were carefully cleaned up and then systematically standardized, which included the standardization of DIGs/disease/target, the classification of DIGs/target, the unification of activity unit, DIGs/target structure, crosslinks to various reference databases and so on. ACDINA has been smoothly running for months and tested from various sites around the world. All ACDINA data can be viewed online and are fully downloadable. To improve the user experience, those downloadable data were categorized to various groups, such as (i) biological activities, (ii) DIG-related general/structural data, (iii) formulation-related information, (iv) API-related general/structural data and (v) DIG's biological target data. Currently, this database can be freely accessed without login requirement by all users at: https://idrblab.org/acdina/.

Data retrieval, processing and preparation for model construction

The prediction of DIG-target interaction was conducted using the biological activities collected to ACDINA. To make the data fit for the construction of prediction model, several steps of data processing were applied. For each DIG-target pair, the DIG was first represented by its structure in the SMILES format [45], and the target was encoded based on its sequence in FASTA format [46]. Then, the DIG and target were digitalized by the *Morgan* fingerprints of 1024 dimensions (with the 'radius' set to 2) generated using *RDKit* package in *Python* [47] and the CTD encoding technique integrated in PROFEAT sever [48, 49], respectively. Third, all DIG-target pairs were divided into interacting and non-interacting datasets according to the biological activity of DIG to its corresponding target. Particularly, each activity was represented based on the popular unit of IC_{50}, EC_{50} or K_i. The DIG-target pairs of high interacting activity (<1 μM, which was widely adopted by existing publication [50] to define the 'active' biological affinity) were grouped into the interacting dataset (also known as the 'positive dataset' and labeled as '1'), while the DIG-target pairs of low interacting activity (>10 μM, which was commonly accepted by the previous study [51] to indicate the 'non-interacting' pairs) were classified into the non-interacting dataset (also known as the 'negative dataset' and labeled as '0'). Fourth, all DIG-target pairs collected from ACDINA were represented by concatenating the DIG's *Morgan* fingerprint to the target's CTD encoding feature, and the Min-Max scaler method [52] was used to normalize the

Downloaded from https://academic.oup.com/bib/article/23/5/bbac160/6582006 by National Science & Technology Library user on 07 October 2022

resulting vectors representing DIG-target pairs (all scaled to [0,1]). Finally, 5-fold cross validation (CV) was applied to assess the performance of the constructed models.

ML techniques for assessing the predictive potential

To assess the predictive potential of DIG-target interaction data collected in ACDINA, a variety of ML techniques were used. These techniques included five classical MLs [including support vector machine (SVM), naive Bayesian (NB), random forest (RF), extreme gradient boosting (XGBoost) and k-nearest neighbors (KNN)] and two deep learnings [such as deep neural networks (DNNs) and convolutional neural networks (CNNs)]. These techniques were integrated into the 5-fold CV, which were implemented in *Python* (version 3.7.11). The model's hyperparameters were carefully tuned using the *optuna* package, and the way to determine these hyperparameters and the resulting optimized parameters was explicitly described in the following section.

Support vector machine

As one of the most widely applied supervised learning methods, SVM model classified samples using hyperplane in high-dimension feature space [53]. Particularly, it determined the decision boundary by finding a hyperplane with the maximum margin and was applied to many aspects of biomedical research [54–64], such as protein function annotation [65], drug–target interaction [66] and drug-like property predictions [67]. Particularly, *sklearn* package offered in the *Python* environment was used to construct model for predicting DIG–target interaction. SVM projected the original feature vectors into a high-dimension space using different types of kernel function, among which the Radial Basis Function was selected in this study to calculate the scalar product between data points that represented two distinct DIG–target interacting/non-interacting pairs. In this study, the hyperparameter C and *gamma* were systematically optimized using *grid search*, and the optimized values equaled to 40 and 0.01 for C and *gamma*, respectively.

Naive Bayesian

The NB classifier was based on the *Bayes*' theorem with independence assumptions between predictors [68]. Although this assumption may be unrealistic for some prediction tasks, NB classifier obtained outstanding performances in pharmaceutical sciences and other relevant directions, even though there were intrinsic attribute dependency for some particular cases. The NB described the conditional probability of an event based on prior knowledge of an event, and calculated the posterior probability of samples from distinct categories and selected the category with high posterior probability as the prediction result. The NB had no complicated hyperparameter and was thus easy to be implemented, which made it popular for different tasks. The NB has been applied to identify new drug and target [69]. In this study, the *sklearn* package was used to develop the NB classifier by assuming a data distribution of multivariate *Bernoulli*. The hyperparameters *alpha* and *binarize* were optimized to 1 and 0.75, respectively.

Extreme gradient boosting

XGBoost is an ensemble learning technique based on gradient boosting and decision tree, which was applied to predict biological activity [70], miRNA-disease association [71] and subcellular location [72]. Particularly, it is optimized model that combines the linear model with a boosting tree one. In this technique, the decision trees are created in sequential form, and the weights are assigned to all independent variables which are then fed into a decision tree. In training process, one tree was built based on the loss generated by the previous weak decision tree–based classifier in each iteration. Finally, the XGBoost achieves stronger learning effect by integrating multiple weak learners [73]. In this study, XGBoost models were built based on *xgboost* package. Those essential hyperparameters *n_estimators*, *learning_rate*, *colsample_bytree*, *max_depth*, *gamma* and *subsample* were optimized to 16, 0.05, 0.4, 3, 0.1 and 0.4, respectively.

Random forest

This technique is decision tree based, which has hierarchical structure and is composed of nodes and branches. It stands out from other tree-based models, and is an ensemble method [74]. This algorithm used a set of mutually independent decision trees to discover the solution to a specific problem. During model construction, the RF grows the trees, and each tree is then trained based on a subset of randomly selected samples. Each decision tree is used to solve the corresponding problem individually, which can produce many classification results. RF ultimately determined the overall solution by considering the majority of the classification votes, which can effectively avoid the issue of overfitting [75]. It has been applied to drug–target interaction prediction [76] and drug discovery [77]. In this work, this method was realized by *ensemble* package in *sklearn*. Three hyperparameters (*n_estimators*, *max_depth* and *max_features*) essential for constructing RF model were optimized to 7, 5, and none, respectively.

K-nearest neighbors

When making a classification decision, KNN only determines the class of the sample according to the nearest sample or several samples in the feature space. The *K* value was a constant defined by the user [78]. Although this technique is regarded as lazy learning, it has been widely adopted in various directions of drug discovery due to its good efficiency and high accuracy. Particularly, it has been applied to drug repositioning and the identification of drug-disease associations [79]. Moreover, it has also been used to construct quantitative structure–activity relationship model for

Downloaded from https://academic.oup.com/bib/article/23/5/bbac160/6582006 by National Science & Technology Library user on 07 October 2022

virtually screening of small molecular inhibitors against G-protein-coupled receptors [80], kinase [81], etc. In this study, *KNeighborsClassifier* implemented in *sklearn* package was applied to build the KNN classifier, and the Euclidean distance was used to measure the distance between any two DIG-target pairs. The hyperparameter *n_neighbors* was optimized to 9.

Deep neural networks

The DNN (also known as multilayer perceptron) is a fully connected neural network with many hidden layers [82]. It realizes the feature learning of input data using nonlinear transformation between simulated neurons, and each layer is composed of various neurons [82]. As a classical deep learning technique, it has been applied to predicting pharmacological property of drug [83] and identifying biomarkers in metabolomics studies [84]. In this study, DNN classifier was built based on the *pytorch* library, which contained three hidden layers with 512, 256 and 64 neurons. Neurons in each hidden layer were activated using a rectified linear unit (ReLU) function. The output layer then generated the classification probability for each DIG-target pair. The training procedure was the well-established backward propagation algorithm implemented using *Adam* optimizer [85]. The hyperparameter of *initial learning rate* was set to 0.001.

Convolutional neural networks

The CNN is a feedforward neural network that uses convolutional and pooling operators, which has been widely applied to the fields of computer vision [86], magnetic resonance imaging [87] and drug discovery [88]. In this study, the *pytorch* library was applied to build the one-dimensional CNN [89], which consisted of seven distinct layers: two convolutional layers, two max-pooling layers, two fully connected layers and one softmax layer. Within each of the convolution layers, 1×5 convolution kernels were used to scan matrix. The first convolutional layer had 64 different convolution kernels and the second one had 128 kernels. After convolution layers, a 1×10 max-pooling layer was then incorporated, and the resulting matrix was then flattened and input into the fully connected layers, which contained two layers of 512 and 64 neurons using the ReLU function. Finally, the softmax layer was adopted to calculate the probability of different categories. CNN model is optimized using *Adam* with a learning rate of 0.001.

Three measurements were used in this study to assess the predictive potential of the DIG-related data collected for the ACDINA, which included accuracy (ACC), area under receiver operating characteristic curve (AUC) and *Matthews* correlation coefficient (MCC). The MCC is considered as one of the most comprehensive metrics due to its collective considerations of both interacting and non-interacting datasets, especially in case of the imbalanced datasets (imbalanced numbers of data in datasets) [90]. MCC = 1 indicates the completely correct classification, and −1 denotes the complete misclassification. For the different models in each round of CV, MCC values were adopted as the key criteria to facilitate the optimization of hyperparameters.

Results and Discussion

Detailed information collected for DIG

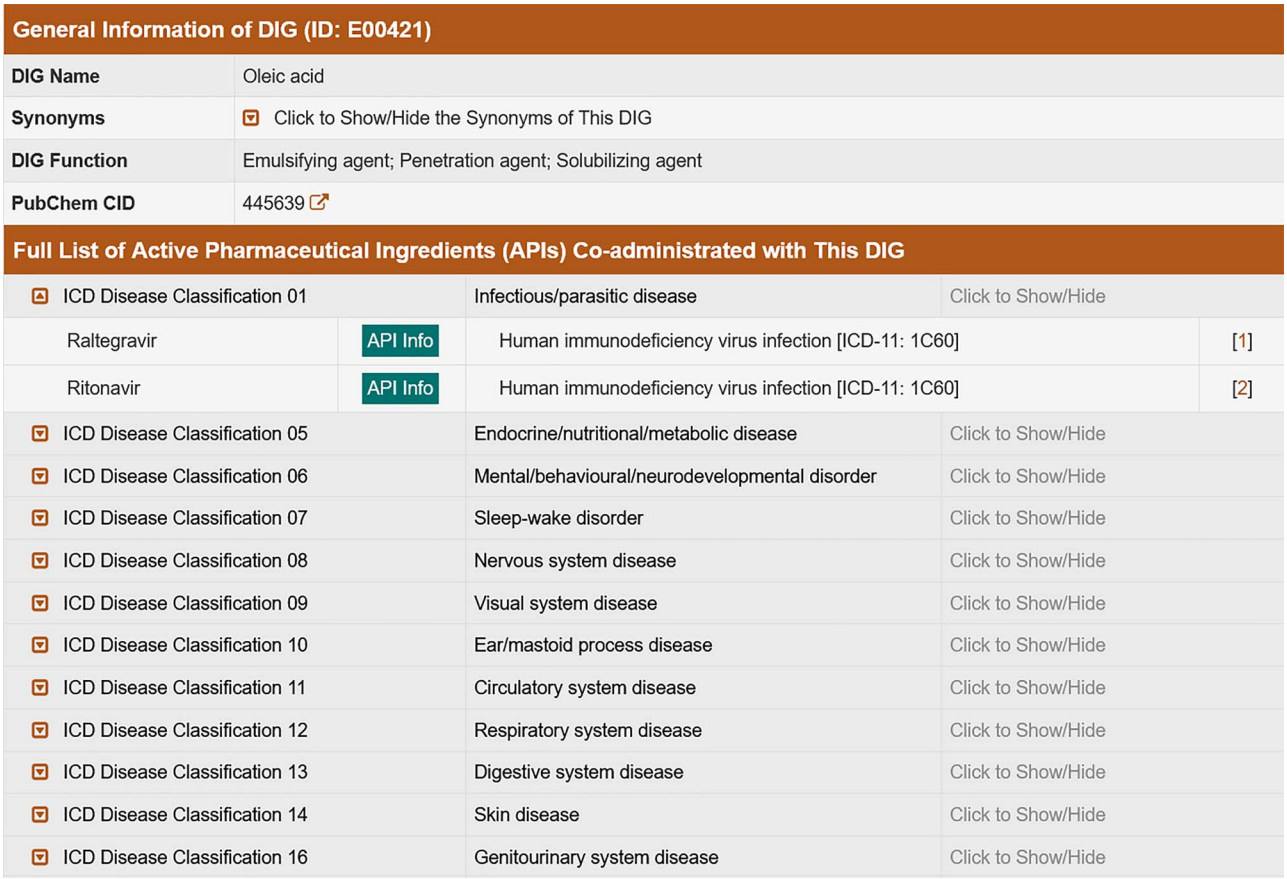
To collect the comprehensive information of the DIGs, multiple steps were applied in this study. First, the full list of drugs approved by US FDA was collected from DrugBank [91] and TTD [45], which led to >2000 FDA-approved drugs. Then, a comprehensive literature review on all these approved drugs was conducted, and a total of 23 949 DFMs that consisted of 711 unique DIGs were identified [19, 28, 92–95]. These DIGs covered the very wide (>50) DIG-functional classes as defined from US Pharmacopeia [37]. Third, all these identified DIGs were manually matched with the compounds in PubChem [96]. Since the name of certain DIG varied when it was applied to different fields, a repository that contained the diverse synonyms for all DIGs was comprehensively collected via a literature review in PubMed and many other existing pharmaceutical databases [19, 22, 25, 96]. Moreover, various molecule representation methods (such as Canonical SMILES and InChI) were provided to facilitate the future estimations/model constructions based on DIG structures. Fourth, various DIG functions in the DFMs (that were critical to formulation design) were carefully identified and systematically recorded based on the well-defined classification system [28]. Some key functions described in ACDINA included antimicrobial preservatives, emulsifying agents, surfactants, buffering agents, etc.

As illustrated in Figure 1 (the page describing a typical DIG oleic acid), the general information of this DIG was provided in the upper section, which included the DIG name, a list of synonyms, DIG functions and the hyperlinks to the other existing pharmaceutical databases. At the bottom section, the full list of APIs co-administrated with this DIG was provided, and these APIs were classified based on the WHO ICD-11 of their corresponding disease indication [97]. For each API, its specific disease indication and references were shown. By clicking the green 'API Info' button, the DIG page (Figure 1) will then be redirected to the page that describes the detailed information of the corresponding API.

Biological targets and activities of DIGs

Since the data of DIG activity are largely dispersed in literature, the PubMed was systematically searched to discover the interaction between DIGs and biological targets. Particularly, keyword combinations of 'interactions + "DIG Name"', 'metabolism + "DIG Name"', 'biological target + "DIG Name"', 'transporters + "DIG Name"', 'adverse reactions + excipient', 'inhibit + drug inactive ingredient', 'inhibit + excipient' and so on were adopted for the literature review, and the resulting publications were manually assessed for retrieving any DIG

Downloaded from https://academic.oup.com/bib/article/23/5/bbac160/6582006 by National Science & Technology Library user on 07 October 2022



General Information of DIG (ID: E00421)			
DIG Name	Oleic acid		
Synonyms	Click to Show/Hide the Synonyms of This DIG		
DIG Function	Emulsifying agent; Penetration agent; Solubilizing agent		
PubChem CID	445639		
Full List of Active Pharmaceutical Ingredients (APIs) Co-administrated with This DIG			
ICD Disease Classification 01		Infectious/parasitic disease	Click to Show/Hide
Raltegravir	API Info	Human immunodeficiency virus infection [ICD-11: 1C60]	[1]
Ritonavir	API Info	Human immunodeficiency virus infection [ICD-11: 1C60]	[2]
ICD Disease Classification 05		Endocrine/nutritional/metabolic disease	Click to Show/Hide
ICD Disease Classification 06		Mental/behavioural/neurodevelopmental disorder	Click to Show/Hide
ICD Disease Classification 07		Sleep-wake disorder	Click to Show/Hide
ICD Disease Classification 08		Nervous system disease	Click to Show/Hide
ICD Disease Classification 09		Visual system disease	Click to Show/Hide
ICD Disease Classification 10		Ear/mastoid process disease	Click to Show/Hide
ICD Disease Classification 11		Circulatory system disease	Click to Show/Hide
ICD Disease Classification 12		Respiratory system disease	Click to Show/Hide
ICD Disease Classification 13		Digestive system disease	Click to Show/Hide
ICD Disease Classification 14		Skin disease	Click to Show/Hide
ICD Disease Classification 16		Genitourinary system disease	Click to Show/Hide

Figure 1. The web page describing a typical DIG named *oleic acid*. The general information of this DIG was provided in the upper section, and the full list of APIs co-administrated with this DIG was provided at the bottom. All APIs were classified by World Health Organization ICD-11 of their corresponding disease indication. For each API, its specific disease indication and references were shown. By clicking the green 'API Info' button, this page will be redirected to a new page showing the data of the corresponding API.

activities-related data. As a result, the collected data included the biological target, the hyperlink to the additional data of this target in other existing pharmaceutical databases (such as UniProt), experimental designs, the tested species in that experiment and DIG's experimentally verified activity to its biological target. The latest ACDINA provided 1318 activity data of 351 DIGs that regulate 362 biological targets from very diverse biochemical families, such as G-protein-coupled receptor, cytochrome P450, transporter and channel. As shown in Figure 2, the distribution of the activity data among all DIGs collected in ACDINA was provided. The vast majority (~85%) of the biological activities were ≥ 1 μM, and over 60% of the activities were within the range between 1 and 100 μM. In other words, although a large number of DIGs demonstrated their activities against some targets, their corresponding interacting values (IC_{50}/Ki) were not so significant (especially when comparing with the interacting values of APIs, which were usually ≤ 1 μM). However, the DIGs can usually achieve much higher concentrations (up to 100 times in gastrointestinal tract) than that reached by API [2]. Therefore, those activities described in Figure 2 raise serious concerns regarding their clinical toxicity [3], DIG–drug interaction [4] and drug resistance [5], which should be considered with caution. Moreover, those experimental consequences of the studied activity shown in ACDINA were also very diverse, which included the regulation of enzyme function, influence of drug efflux, alteration of target expressions and so on. All DIG activities can be accessed and retrieved using the user-friendly search strategies shown in ACDINA homepage, DIG subpage (by clicking the 'Drug Inactive Ingredient' button in navigation bar) and target subpage (by clicking the 'Biological Target of DIG' button).

As illustrated in Figure 3 (the page showing the DIG benzoic acid), the general information of this DIG was provided in the upper section (similar to Figure 1). At the bottom section, the full list of biological targets (DBTs) regulated by this DIG was provided, and all these targets were classified based on their biochemical family. For the particular example illustrated in Figure 3, there were diverse families covered by the DBTs of this DIG, which included G-protein-coupled receptors, oxidoreductase, primary active transporter, lyase/isomerase/ligase and so on. Within the family of primary active transporter, there is a DBT named 'Bile salt export pump'. The DIG-related activity data were explicitly provided at the bottom (biological target, hyperlink to the additional data of this target in other existing databases, experimentally tested species and DIG's experimentally verified activity

Downloaded from https://academic.oup.com/bib/article/23/5/bbac160/6582006 by National Science & Technology Library user on 07 October 2022

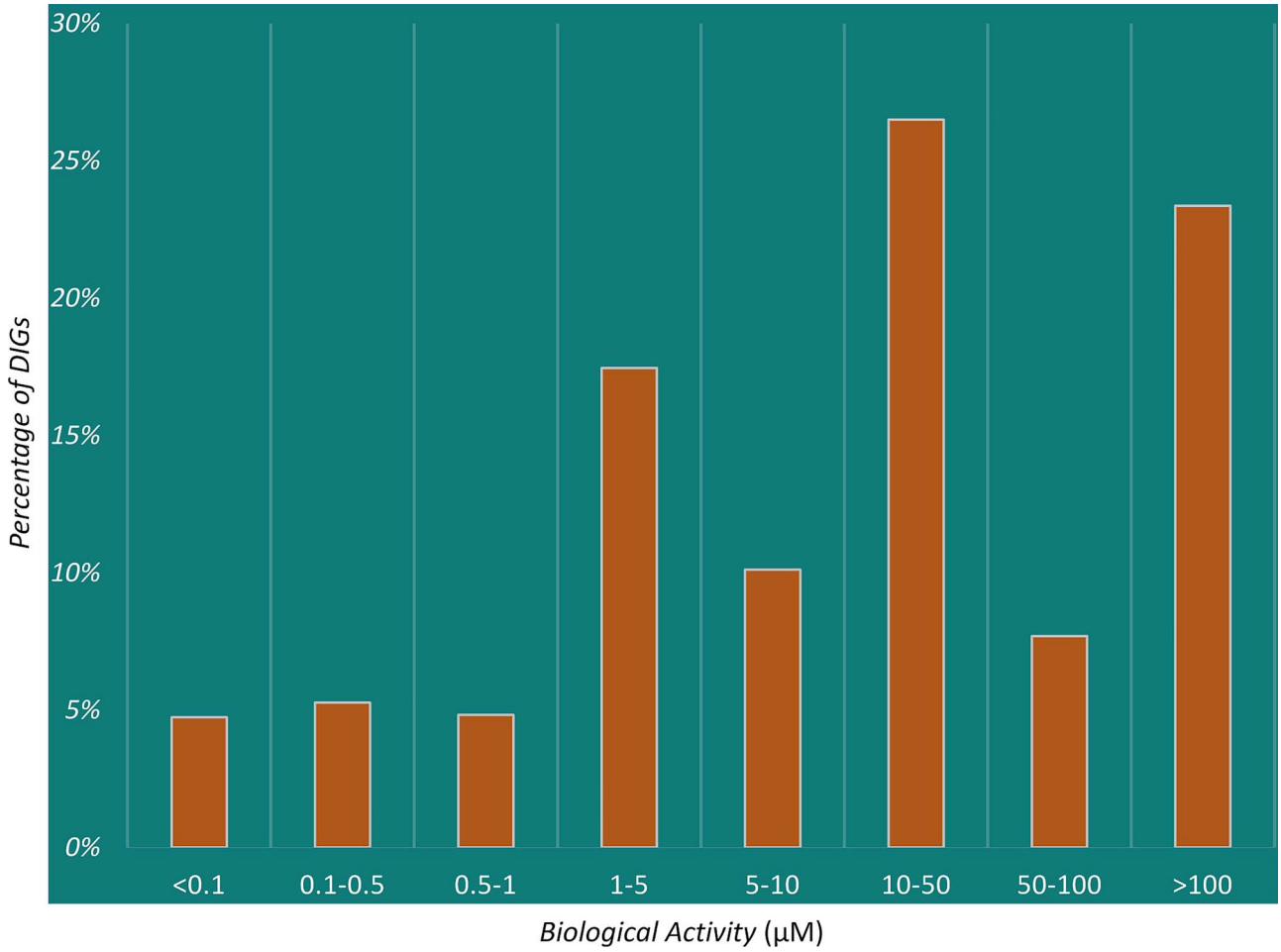


Figure 2. The distribution of the activity data among all DIGs collected in ACDINA. The vast majority (~85%) of the biological activities were ≥1 μM, and over 60% of the activity data were in the range between 1 and 100 μM. In other words, although many DIGs showed their activities against biological target, the corresponding interacting values (IC50/Ki) were not so significant (especially when comparing with the interacting values of the APIs, which were usually ≤1 μM).

data to its biological target). By clicking the orange 'DBT Info' button, this page (Figure 3) will be redirected to a new DBT page, which further provided the information of DBT synonyms, biochemical family, tested organism, gene name, etc.

Comprehensive description on DFMs approved by FDA

The approved DFMs were collected from the official website of the US FDA [21]. Moreover, some popular databases (such as Pillbox and Drug Daily) were reviewed, from which thousands of oral DFMs used in the market were collected. Based on these collected data and additional literature reviews in PubMed, a total of 23 949 DFMs approved by FDA were collected, which covered over 2000 approved APIs. Particularly, these DFMs are distributed in very wide range of dosage forms, and some of the popular forms include extended-release tablet (5136 formulations), extended-release capsule (1881 formulations), delayed-release tablet (962 formulations), chewable tablet (617 formulations), delayed-release capsule (542 formulations) and solution (227 formulations); these DFMs also belong to the very diverse routes of administration, and some top-ranked routes include oral (23 529 formulations), intravenous (386 formulations), subcutaneous (158 formulations), intramuscular (90 formulations), topical (97 formulations), ophthalmic (52 formulations), nasal (15 formulations), inhalation (12 formulations) and so on. Moreover, some of the new drug delivery systems emerged in recent year have also been included in the database. One of the typical examples is the liposome [98], which is a lipid bilayer-based spherical vesicle used as drug delivery system for the administration of nutrients and drugs. With the increasingly accumulated biological data for the new systems, the impacts of ACDINA on the future practice of drug discovery and precision medicine would be further extended in the near future.

As illustrated in Figure 4 (the page showing the API ranitidine), the general information of this API was provided in the upper section, which included the API name, synonyms, clinical status, approved disease indications with ICD-11 classification, and the hyperlinks to the other existing pharmaceutical databases. At the bottom, the full list of drug formulations (DFMs) that contain this API is provided, and all these formulations are

Downloaded from https://academic.oup.com/bib/article/23/5/bbac160/6582006 by National Science & Technology Library user on 07 October 2022

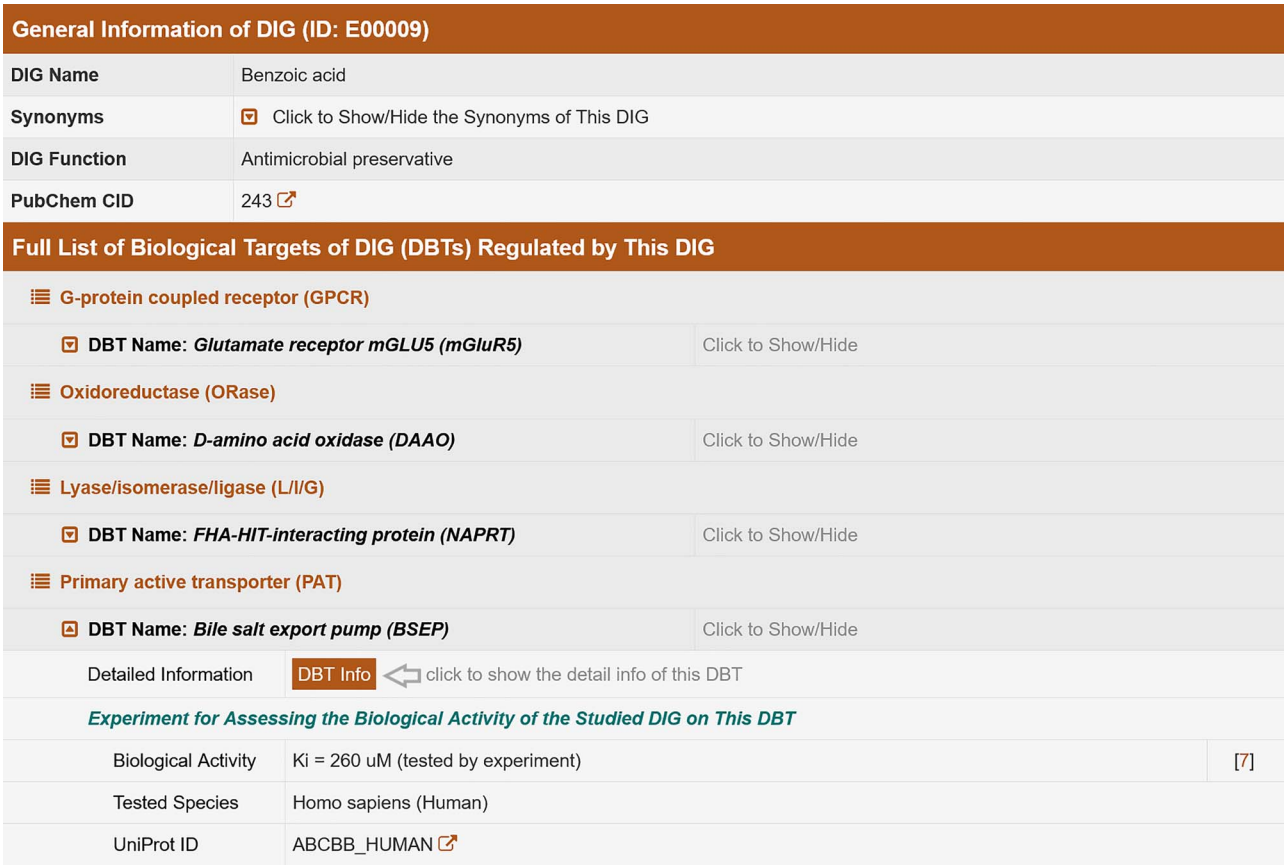


Figure 3. The web page describing a typical DIG named *benzoic acid*. The general information of this DIG was provided in the upper section, and full list of biological targets (DBTs) regulated by this DIG was provided at the bottom. All targets were classified by their biochemical families. DIG-related activity data were shown under the texts in green, which included biological target, hyperlink to the additional data of this target in other available databases, experimentally tested species and DIG's experimentally verified activity to its target. By clicking 'DBT Info' button, this page will be redirected to a new page, which shows the data of DBT synonyms, biochemical family, tested organism, gene name and so on.

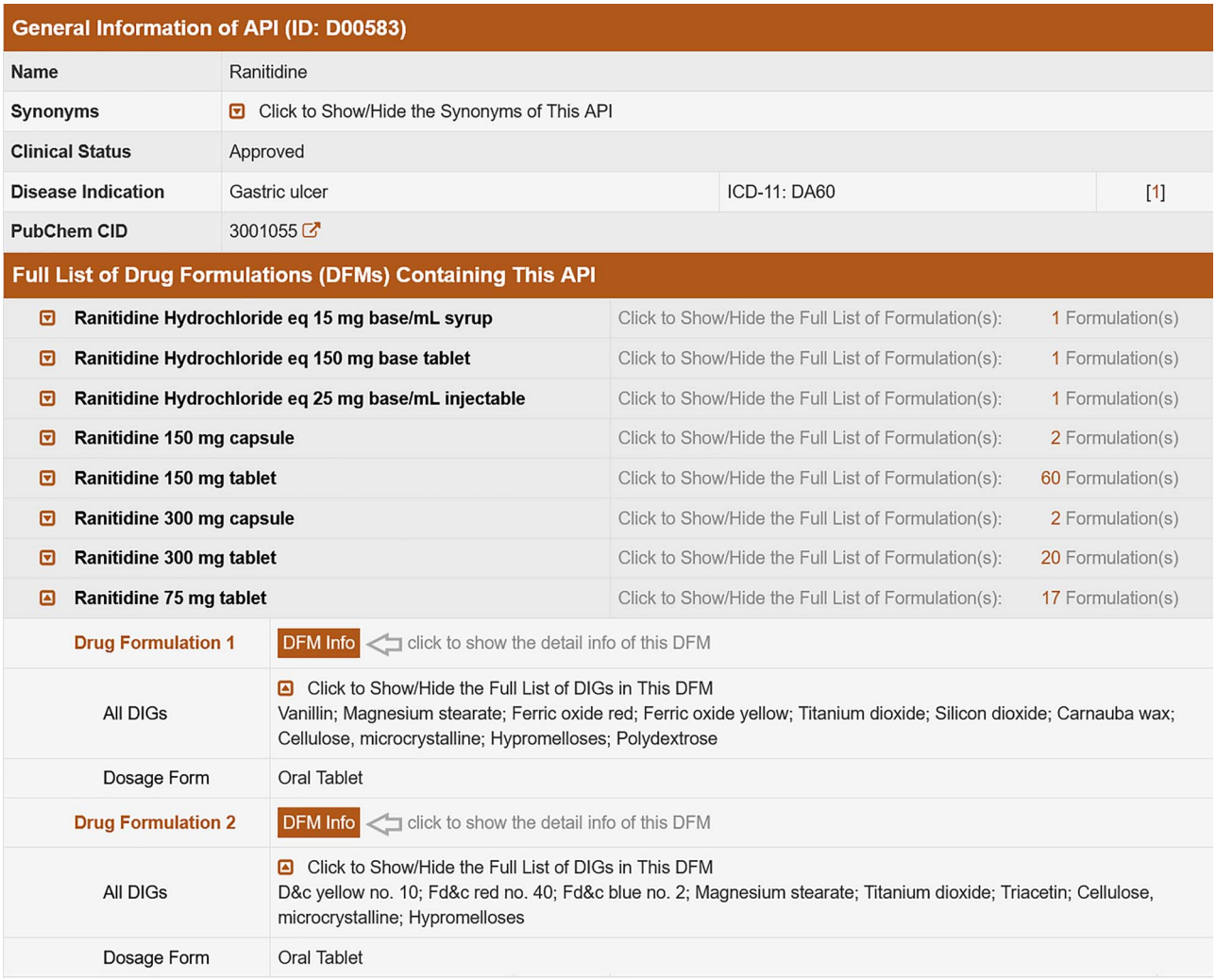
grouped based on their dosage, dosage form and route of administration. Within each formulation, all DIGs and corresponding dosage forms were explicitly described. By clicking the orange 'DFM Info' button, current API page (Figure 4) will be redirected to a new DFM page which is illustrated in Figure 5.

As shown in Figure 5 (a page of DFM ranitidine 75 mg tablet), the general information of this DFM was provided in the upper section, which included the DFM name, developers/companies, API and the full list of DIGs. At the bottom, the full list of the biological targets (DBTs) regulated by the DIGs in this DFM was provided, and all these DBTs were grouped using their biochemical families. For the particular example shown in Figure 5, there were some families covered by the DBTs of this DFM, which included G-protein-coupled receptors, oxidoreductase, transferase, hydrolase and so on. Within the family of G-protein-coupled receptors, there is a DBT named 'adenosine receptor A3'. The DIG-related activity data were explicitly provided at the bottom (biological target, hyperlink to the additional data of this target in other existing databases, experimentally tested species and DIG's experimentally verified activity data to its biological target). By clicking the orange 'DBT Info' button, this page (Figure 5) will be redirected to a new DBT page, which further provided the information of DBT synonyms, biochemical family, tested organism, gene name, etc.

Assessing the predictive potential of ACDINA using ML

The application of AI technology to a specific research and development direction has emerged to be very popular, which has attracted great interest from numerous research fields [99]. In the field of predicting DIG-target interactions, the newly developed ACDINA database in this study should therefore be considered as a comprehensive and first-hand knowledge base to meet the urgent demands of the related research community. To have a preliminary understanding of the predictive potential of the DIG's biological activity data in ACDINA, the ML techniques that have achieved impressive performances in many other directions were adopted in this study to enable an in-depth assessment. There were seven techniques applied in this study, which were five classical MLs (including SVM, NB, RF, XGBoost and KNN) and two deep learnings (including DNNs and CNNs). If the good prediction performance

Downloaded from https://academic.oup.com/bib/article/23/5/bbac160/6582006 by National Science & Technology Library user on 07 October 2022



General Information of API (ID: D00583)

Name	Ranitidine		
Synonyms	Click to Show/Hide the Synonyms of This API		
Clinical Status	Approved		
Disease Indication	Gastric ulcer	ICD-11: DA60	[1]
PubChem CID	3001055		

Full List of Drug Formulations (DFMs) Containing This API

Ranitidine Hydrochloride eq 15 mg base/mL syrup	Click to Show/Hide the Full List of Formulation(s):	1 Formulation(s)
Ranitidine Hydrochloride eq 150 mg base tablet	Click to Show/Hide the Full List of Formulation(s):	1 Formulation(s)
Ranitidine Hydrochloride eq 25 mg base/mL injectable	Click to Show/Hide the Full List of Formulation(s):	1 Formulation(s)
Ranitidine 150 mg capsule	Click to Show/Hide the Full List of Formulation(s):	2 Formulation(s)
Ranitidine 150 mg tablet	Click to Show/Hide the Full List of Formulation(s):	60 Formulation(s)
Ranitidine 300 mg capsule	Click to Show/Hide the Full List of Formulation(s):	2 Formulation(s)
Ranitidine 300 mg tablet	Click to Show/Hide the Full List of Formulation(s):	20 Formulation(s)
Ranitidine 75 mg tablet	Click to Show/Hide the Full List of Formulation(s):	17 Formulation(s)

Drug Formulation 1	DFM Info click to show the detail info of this DFM
All DIGs	Click to Show/Hide the Full List of DIGs in This DFM Vanillin; Magnesium stearate; Ferric oxide red; Ferric oxide yellow; Titanium dioxide; Silicon dioxide; Carnauba wax; Cellulose, microcrystalline; Hypromelloses; Polydextrose
Dosage Form	Oral Tablet
Drug Formulation 2	DFM Info click to show the detail info of this DFM
All DIGs	Click to Show/Hide the Full List of DIGs in This DFM D&c yellow no. 10; Fd&c red no. 40; Fd&c blue no. 2; Magnesium stearate; Titanium dioxide; Triacetin; Cellulose, microcrystalline; Hypromelloses
Dosage Form	Oral Tablet

Figure 4. The web page describing a typical API named *ranitidine*. The general information of this API was provided in the upper section, and full list of DFMs containing this API was provided at the bottom. All DFMs were grouped based on their dosage, dosage form and route of administration. Within each formulation, all DIGs and corresponding dosage forms were explicitly described. By clicking the orange 'DFM Info' button, current API page will be redirected to a new page illustrating the corresponding formulation.

could be obtained from these assessments, it would be reasonable to expect an improved prediction result based on ACDINA data by the application of more advanced and sophisticated Artificial Intelligence techniques. The detailed descriptions on each ML technique together with the adopted performance assessment metrics were explicitly shown in 'ML Techniques for Assessing the Predictive Potential' section of Materials and Methods, and the assessment results were provided in Table 1.

ACC and AUC were the popular metrics in performance assessment. The ranges of these metric values were from 0 to 1. As reported, a value (ACC/AUC) greater than 0.7 was commonly used as an indicator of a good prediction [100]. As shown in Table 1, DNN gave the best ACC (0.8564) and AUC (0.8331) among all seven ML techniques (0.8564). The ACCs of all seven techniques and the AUCs of six techniques (except RF) illustrated good prediction results (with the values of ACC/AUC over 0.7). In other words, these results showed that there were great potentials in ACDINA data to facilitate the prediction of new DIG-target interaction.

To assess the effect of imbalanced data on models' prediction performance, the well-established metric MCC was calculated and compared among different ML techniques. As demonstrated in Table 1, DNN gave the best MCC among all techniques (0.4610), while MCC of XGBoost was the lowest (0.3002). On the one hand, the variation among the MCC values of these techniques was not so significant (all lower than 0.5), which indicated that the prediction potentials of the ACDINA data among seven MLs might be equivalent. On the other hand, the MCC was within [−1,1] ('1' means completely correct classification; '−1' denotes complete misclassification; '>0' indicates that the model is better than random prediction), and the MCC values of all seven MLs could thus be considered as *fair*. In other words, there were great rooms for improvement (from <0.5 to 1), which inspired us to conduct research to further improve the

Downloaded from https://academic.oup.com/bib/article/23/5/bbac160/6582006 by National Science & Technology Library user on 07 October 2022

General Information of DFM (ID: F18866)

Name	Ranitidine 75 mg tablet			
Company	Family Dollar; Harris Teeter; Meijer; Wal-Mart Stores			
Active Pharmaceutical Ingredient (API)	Ranitidine	API Info ⇦ click to show the detail info of this API		
	Gastric ulcer		ICD-11: DA60	Approved
Drug Inactive Ingredient (DIGs)	**DIG ID**	**DIG Info**	**DIG Name**	**DIG Functional Class**
	E00338	DIG Info	Allura red AC dye	Colorant
	E00446	DIG Info	FD&C blue no. 2	Colorant
	E00208	DIG Info	Magnesium stearate	lubricant
	E00309	DIG Info	Quinoline yellow WS	Colorant
	E00322	DIG Info	Titanium dioxide	Coating agent ...
	E00080	DIG Info	Triacetin	Humectant ...

Full List of Biological Targets of DIG (DBTs) Regulated by DIG(s) in This DFM

- **G-protein coupled receptor (GPCR)**
  - **DBT Name: *Adenosine receptor A3 (AA3R)*** — Click to Show/Hide
    - Detailed Information: DBT Info ⇦ click to show the detail info of this DBT
    - ***Experiment for Assessing the Biological Activity of the Studied DIG on This DBT***
      - Interacting DIG: FD&C blue no. 2
      - Biological Activity: IC50 = 1 uM (tested by experiment) [1]
      - Tested Species: Homo sapiens (Human)
      - UniProt ID: AA3R_HUMAN
  - **DBT Name: *Alpha-2A adrenoceptor (ADA2A)*** — Click to Show/Hide
  - **DBT Name: *Dopamine D3 receptor (DRD3)*** — Click to Show/Hide
- **Oxidoreductase (ORase)**
  - **DBT Name: *Monoamine oxidase A (MAO-A)*** — Click to Show/Hide
  - **DBT Name: *Prostaglandin G/H synthase 2 (COX-2)*** — Click to Show/Hide
- **Transferase (TFase)**
  - **DBT Name: *VEGF-2 receptor (KDR)*** — Click to Show/Hide
- **Hydrolase (HDase)**
  - **DBT Name: *Acetylcholinesterase (ACHE)*** — Click to Show/Hide

Figure 5. A web page demonstrating a typical DFM named *ranitidine 75 mg tablet*. The general information of this DFM was provided in the upper section, and the full list of biological targets (DBTs) regulated by the DIGs in this DFM was provided at the bottom. All DBTs were grouped based on their biochemical family. The DIG-related activity data were provided under the green text (biological target, hyperlinks to the additional data of this target in other existing databases, experimentally tested species, DIG's experimentally verified activity data to its biological target, etc.). By clicking the 'DBT Info' button, this page will be redirected to another page providing the information of DBT synonyms, biochemical family, tested organism, gene name, etc.

prediction performances by considering the imbalance issue that was assessed using MCC.

In this study, a preliminary evaluation on the possibility to elevate the MCC was also conducted. Seven methods that collectively considered all ML techniques were proposed and evaluated in Figure 6. These methods included ANY-M: the DIG-target pairs predicted as interacting by any M ($M = 1 \sim 6$) of the seven ML techniques were considered as interacting; ALL-7: the DIG-target pairs predicted as interacting by all ML techniques were considered as interacting. As described in Table 1, there was no significant improvement between those ACCs of

Downloaded from https://academic.oup.com/bib/article/23/5/bbac160/6582006 by National Science & Technology Library user on 07 October 2022

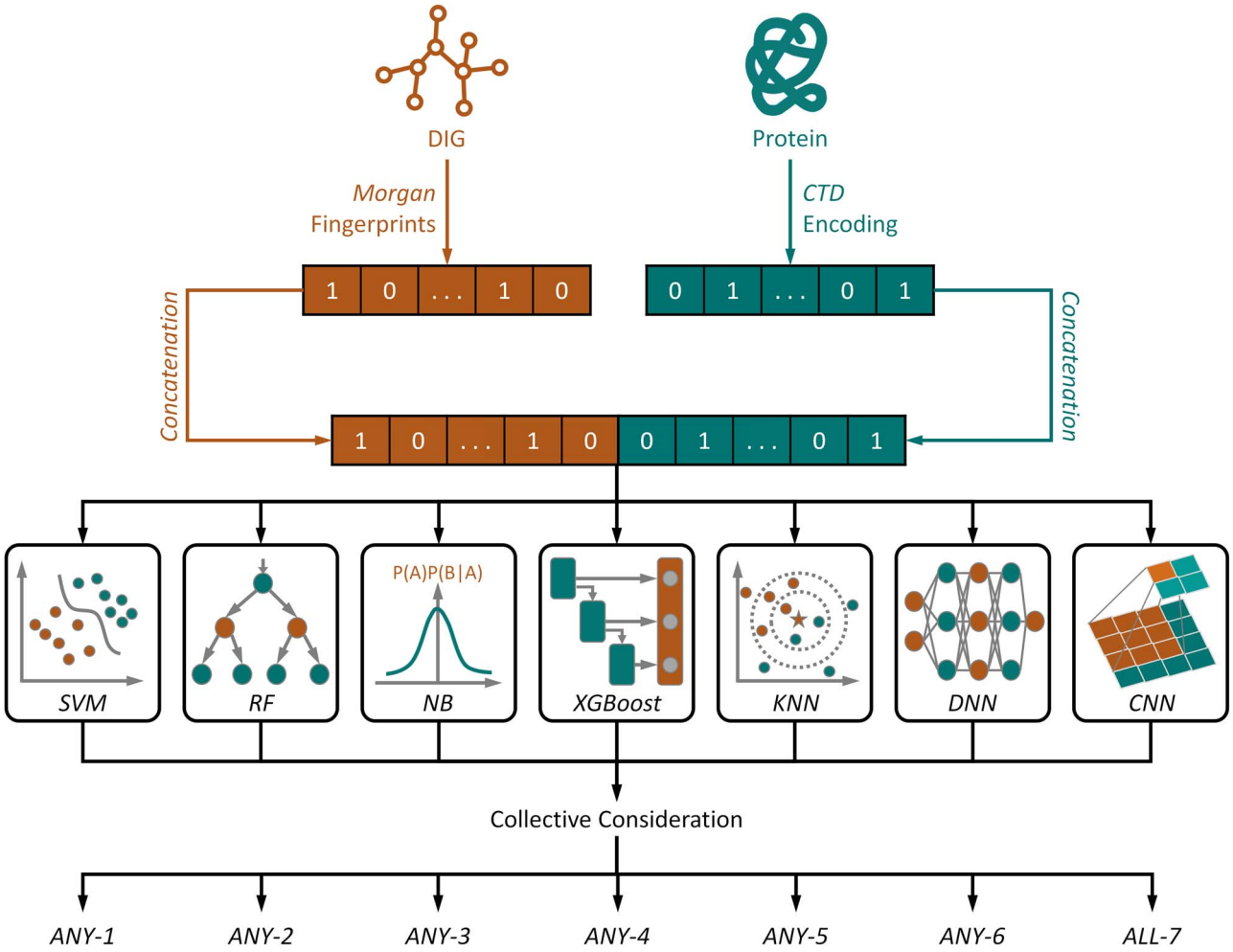


Figure 6. Schematic flow chart of the applications of seven ML techniques and seven methods collectively considering multiple MLs for assessing the predictive potential of DIG activity data collected in ACDINA. These seven techniques included five classical MLs (including SVM, RF, NB, XGBoost and KNN) and two deep learnings (such as DNNs and CNNs). ANY-*M*: DIG-target pairs predicted as interacting by any *M* ($M = 1 \sim 6$) of the seven ML techniques were considered as interacting; ALL-7: the DIG-target pairs predicted as interacting by all seven ML techniques were considered as interacting.

Table 1. The performances of seven single ML techniques and seven methods collectively considering multiple ML techniques. These techniques included five classical MLs (including SVM, NB, RF, XGBoost and KNN) and two deep learnings (such as DNNs and CNNs). ANY-*M*: DIG-target pairs predicted as interacting by any *M* ($M = 1 \sim 6$) of the seven ML techniques were considered as interacting; ALL-7: the DIG-target pairs predicted as interacting by all seven ML techniques were considered as interacting. All models were constructed and assessed based on 5-fold CV, and the performances reported below were the average values among five CVs

Technique/Method		ACC	AUC	MCC
Single Technique	*SVM*	0.8465	0.7441	0.3469
	NB	0.8174	0.8143	0.3897
	RF	0.8365	0.6985	0.3181
	XGBoost	0.8398	0.7509	0.3002
	KNN	0.8548	0.8212	0.4195
	DNN	0.8564	0.8331	0.4610
	CNN	0.8390	0.8142	0.4202
Collective Method	*ANY-1*	0.8224	–	0.5635
	ANY-2	0.8515	–	0.5298
	ANY-3	0.8581	–	0.4820
	ANY-4	0.8465	–	0.3835
	ANY-5	0.8382	–	0.2932
	ANY-6	0.8324	–	0.2049
	ALL-7	0.8266	–	0.1744

Downloaded from https://academic.oup.com/bib/article/23/5/bbac160/6582006 by National Science & Technology Library user on 07 October 2022

newly proposed seven Collective Methods and that of the Single Techniques. However, the MCCs of both ANY-1 and ANY-2 were substantially increased comparing with those of seven single techniques (all lower than 0.5). These results showed that the prediction performance assessed using MCCs could be further improved by optimizing the architecture of ML techniques, and significant improvement could be achieved if the advanced AI techniques were applied in the future. All in all, based on the assessment using different metrics above, the data of DIG's biological activities in ACDINA were expected to have great potential to facilitate the prediction of novel DIG-target interactions and other research directions in pharmaceutical sciences.

Conclusion

This work (i) described the largest number of well-defined DIGs and DFMs that were manually curated and carefully confirmed using all FDA-approved drugs; (ii) provided for the first time the comprehensive biological activities for both DIGs and DFMs, and all data were available and downloadable online; and (iii) fully referenced the biological targets of each DIG and formulation by cross-linking them to available databases that provide their pharmaceutical/biological characteristics. Moreover, various popular ML techniques were used to assess the predictive potential of the collected data, which was the first assessment on the possibility to predict DIG's biological activities. As the activity data of DIG are critical for current pharmaceutical research, this study is expected to give significant implication for the future practice of molecular interaction [101–104], drug discovery [105–111] and precision medicine [112–115].

Key Points

- This study described the largest number of well-defined DIGs and drug formulations (DFMs) that were manually curated and carefully confirmed using all FDA-approved drugs.
- This study provided, for the first time, the comprehensive biological activity data for both DIGs and DFMs, and all data were available and downloadable online.
- The study fully referenced the targets of each DIG and formulation by cross-linking them to available databases that describe their pharmaceutical/biological characteristics.
- Machine learning was used to assess the predictive potential of the collected data, which was the first assessment on the possibility to predict the DIG's biological activity.

Data availability statement

All data in the manuscript are collected and available in ACDINA database.

Funding

Natural Science Foundation of Zhejiang Province (LR21-H300001); National Natural Science Foundation of China (81872798, U1909208); Leading Talent of the 'Ten Thousand Plan' – National High-Level Talents Special Support Plan of China; 'Double Top-Class' University Project (181201∗194232101); Key R&D Program of Zhejiang Province (2020C03010); Fundamental Research Fund for Central Universities (2018QNA7023); Westlake Laboratory (Westlake Laboratory of Life Sciences and Biomedicine); Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare; Alibaba Cloud; Information Technology Center of Zhejiang University.

References

1. Reker D, Blum SM, Steiger C, *et al.* Inactive ingredients in oral medications. *Sci Transl Med* 2019;**11**:eaau6753.
2. Zou L, Spanogiannopoulos P, Pieper LM, *et al.* Bacterial metabolism rescues the inhibition of intestinal drug absorption by food and drug additives. *Proc Natl Acad Sci U S A* 2020;**117**:16009–18.
3. Turner MA, Duncan JC, Shah U, *et al.* Risk assessment of neonatal excipient exposure: lessons from food safety and other areas. *Adv Drug Deliv Rev* 2014;**73**:89–101.
4. Reker D, Shi Y, Kirtane AR, *et al.* Machine learning uncovers food- and excipient-drug interactions. *Cell Rep* 2020;**30**:3710–6.
5. Sosnik A. Reversal of multidrug resistance by the inhibition of ATP-binding cassette pumps employing generally recognized as safe (GRAS) nanopharmaceuticals: a review. *Adv Drug Deliv Rev* 2013;**65**:1828–51.
6. Gopinath K, Karthikeyan M. Understanding the evolutionary relationship of M2 channel protein of influenza a virus and its structural variation and drug resistance. *Curr Bioinform* 2017;**12**:265–74.
7. Robson B. COVID-19 coronavirus spike protein analysis for synthetic vaccines, a peptidomimetic antagonist, and therapeutic drugs, and analysis of a proposed achilles' heel conserved region to minimize probability of escape mutations and drug resistance. *Comput Biol Med* 2020;**121**:103749.
8. Fu T, Li F, Zhang Y, *et al.* VARIDT 2.0: structural variability of drug transporter. *Nucleic Acids Res* 2022;**50**:D1417–31.
9. Pottel J, Armstrong D, Zou L, *et al.* The activities of drug inactive ingredients on biological targets. *Science* 2020;**369**:403–13.
10. Mai Y, Ashiru-Oredope DAI, Yao Z, *et al.* Boosting drug bioavailability in men but not women through the action of an excipient. *Int J Pharm* 2020;**587**:119678.
11. Xiao L, Yi T, Chen M, *et al.* A new mechanism for increasing the oral bioavailability of scutellarin with cremophor EL: activation of MRP3 with concurrent inhibition of MRP2 and BCRP. *Eur J Pharm Sci* 2016;**93**:456–67.
12. Yang L, Jiao X. Distinguishing enzymes and non-enzymes based on structural information with an alignment free approach. *Curr Bioinform* 2021;**16**:44–52.
13. Kumari M, Subbarao N. Deep learning model for virtual screening of novel 3C-like protease enzyme inhibitors against SARS coronavirus diseases. *Comput Biol Med* 2021;**132**:104317.
14. Dorier M, Brun E, Veronesi G, *et al.* Impact of anatase and rutile titanium dioxide nanoparticles on uptake carriers and efflux pumps in caco-2 gut epithelial cells. *Nanoscale* 2015;**7**:7352–60.

Downloaded from https://academic.oup.com/bib/article/23/5/bbac160/6582006 by National Science & Technology Library user on 07 October 2022

15. Gu YZ, Chu X, Houle R, *et al.* Polyethlyene glycol 200 can protect rats against drug-induced kidney toxicity through inhibition of the renal organic anion transporter 3. *Toxicol Sci* 2019;**172**:155–66.
16. Chen L, Zhang YH, Zou Q, *et al.* Analysis of the chemical toxicity effects using the enrichment of gene ontology terms and KEGG pathways. *Biochim Biophys Acta* 2016;**1860**:2619–26.
17. Elbadawi M, Gaisford S, Basit AW. Advanced machine-learning techniques in drug discovery. *Drug Discov Today* 2021;**26**:769–77.
18. Wood VE, Groves K, Cryar A, *et al.* HDX and in silico docking reveal that excipients stabilize G-CSF via a combination of preferential exclusion and specific hotspot interactions. *Mol Pharm* 2020;**17**:4637–51.
19. USF. FDA IIG: inactive ingredient search for approved drug products. *Official Website US FDA* 2021.
20. Salunke S, Brandys B, Giacoia G, *et al.* The STEP (safety and toxicity of excipients for paediatrics) database: part 2 - the pilot version. *Int J Pharm* 2013;**457**:310–22.
21. USF. Drugs@FDA: FDA-approved drugs. In: *Official Website US FDA*, 2021.
22. US N. Pillbox of the national library of medicine (NLM). In: *Official Website US NIH*, 2009.
23. Irwin JJ, Pottel J, Zou L, *et al.* A molecular basis for innovation in drug excipients. *Clin Pharmacol Ther* 2017;**101**:320–3.
24. Berzins K, Sales RE, Barnsley JE, *et al.* Low-wavenumber Raman spectral database of pharmaceutical excipients. *Vib Spectrosc* 2020;**107**:103021.
25. Mendez D, Gaulton A, Bento AP, *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res* 2019;**47**:D930–40.
26. Sayers EW, Beck J, Bolton EE, *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res* 2021;**49**:D10–7.
27. Gilson MK, Liu T, Baitaluk M, *et al.* BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res* 2016;**44**:D1045–53.
28. Rowe RC, Sheskey PJ, Quinn ME. *Handbook of Pharmaceutical Excipients*. London: Pharm Press, 2009.
29. Li YH, Li XX, Hong JJ, *et al.* Clinical trials, progression-speed differentiating features and swiftness rule of the innovative targets of first-in-class drugs. *Brief Bioinform* 2020;**21**:649–62.
30. Yang H, Qin C, Li YH, *et al.* Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Res* 2016;**44**:D1069–74.
31. Fu T, Zheng G, Tu G, *et al.* Exploring the binding mechanism of metabotropic glutamate receptor 5 negative allosteric modulators in clinical trials by molecular dynamics simulations. *ACS Chem Nerosci* 2018;**9**:1492–502.
32. Zhu F, Qin C, Tao L, *et al.* Clustered patterns of species origins of nature-derived drugs and clues for future bioprospecting. *Proc Natl Acad Sci U S A* 2011;**108**:12943–8.
33. Zhu F, Shi Z, Qin C, *et al.* Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res* 2012;**40**:D1128–36.
34. Zhang Y, Zheng QC. In silico analysis revealed a unique binding but ineffective mode of amantadine to influenza virus B M2 channel. *J Phys Chem Lett* 2021;**12**:1169–74.
35. Zhang Y, Zhang HX, Zheng QC. In silico study of membrane lipid composition regulating conformation and hydration of influenza virus B M2 channel. *J Chem Inf Model* 2020;**60**:3603–15.
36. Lin B, Zhang H, Zheng Q. How do mutations affect the structural characteristics and substrate binding of CYP21A2? An investigation by molecular dynamics simulations. *Phys Chem Chem Phys* 2020;**22**:8870–7.
37. Piervincenzi RT. The pharmacopeia of the United States of America 4th decennial revision. *Natl Pharmacopoeia* 2012;**1059**:599–608.
38. UniProt C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 2021;**49**:D480–9.
39. Li YH, Yu CY, Li XX, *et al.* Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Res* 2018;**46**:D1121–7.
40. Yin J, Sun W, Li F, *et al.* VARIDT 1.0: variability of drug transporter database. *Nucleic Acids Res* 2020;**48**:D1042–50.
41. Bairoch A. The cellosaurus, a cell-line knowledge resource. *J Biomol Tech* 2018;**29**:25–38.
42. Yin J, Li F, Zhou Y, *et al.* INTEDE: interactome of drug-metabolizing enzymes. *Nucleic Acids Res* 2021;**49**:D1233–43.
43. Wang YL, Wang F, Shi XX, *et al.* Cloud 3D-QSAR: a web tool for the development of quantitative structure-activity relationship models in drug discovery. *Brief Bioinform* 2021;**22**:bbaa276.
44. Hao GF, Jiang W, Ye YN, *et al.* ACFIS: a web server for fragment-based drug discovery. *Nucleic Acids Res* 2016;**44**:W550–6.
45. Zhou Y, Zhang Y, Lian X, *et al.* Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents. *Nucleic Acids Res* 2022;**50**:D1398–407.
46. Hong J, Luo Y, Mou M, *et al.* Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery. *Brief Bioinform* 2020;**21**:1825–36.
47. Bento AP, Hersey A, Felix E, *et al.* An open source chemical structure curation pipeline using RDKit. *J Chem* 2020;**12**:51.
48. Hong J, Luo Y, Zhang Y, *et al.* Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning. *Brief Bioinform* 2020;**21**:1437–47.
49. Rao HB, Zhu F, Yang GB, *et al.* Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res* 2011;**39**:W385–90.
50. Li X, Li Z, Wu X, *et al.* Deep learning enhancing kinome-wide polypharmacology profiling: model construction and experiment validation. *J Med Chem* 2020;**63**:8723–37.
51. Avram S, Bora A, Halip L, *et al.* Modeling kinase inhibition using highly confident data sets. *J Chem Inf Model* 2018;**58**:957–67.
52. Li B, Tang J, Yang Q, *et al.* NOREVA: normalization and evaluation of MS-based metabolomics data. *Nucleic Acids Res* 2017;**45**:W162–70.
53. Lister AP, Highmore CJ, Hanrahan N, *et al.* Multi-excitation Raman spectroscopy for label-free, strain-level characterization of bacterial pathogens in artificial sputum media. *Anal Chem* 2022;**94**:669–77.
54. Li D, Ju Y, Zou Q. Protein folds prediction with hierarchical structured SVM. *Curr Proteomics* 2016;**13**:79–85.
55. Fu J, Zhang Y, Wang Y, *et al.* Optimization of metabolomic data processing using NOREVA. *Nat Protoc* 2022;**17**:129–51.
56. Li F, Zhou Y, Zhang Y, *et al.* POSREG: proteomic signature discovered by simultaneously optimizing its reproducibility and generalizability. *Brief Bioinform* 2022;**23**:bbac040.
57. Fu J, Zhang Y, Liu J, *et al.* Pharmacometabonomics: data processing and statistical analysis. *Brief Bioinform* 2021;**22**:bbab138.

Downloaded from https://academic.oup.com/bib/article/23/5/bbac160/6582006 by National Science & Technology Library user on 07 October 2022

58. Tang J, Mou M, Wang Y, *et al.* MetaFS: performance assessment of biomarker discovery in metaproteomics. *Brief Bioinform* 2021;**22**:bbaa105.
59. Yang Q, Li B, Tang J, *et al.* Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data. *Brief Bioinform* 2020;**21**:1058–68.
60. Zhu F, Li XX, Yang SY, *et al.* Clinical success of drug targets prospectively predicted by in silico study. *Trends Pharmacol Sci* 2018;**39**:229–31.
61. Yang Q, Wang Y, Zhang Y, *et al.* NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data. *Nucleic Acids Res* 2020;**48**:W436–48.
62. Tang J, Fu J, Wang Y, *et al.* ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. *Brief Bioinform* 2020;**21**:621–36.
63. Li F, Zhou Y, Zhang X, *et al.* SSizer: determining the sample sufficiency for comparative biological study. *J Mol Biol* 2020;**432**: 3411–21.
64. Li YH, Xu JY, Tao L, *et al.* SVM-Prot 2016: a web-server for machine learning prediction of protein functional families from sequence irrespective of similarity. *PLoS One* 2016;**11**: e0155290.
65. Han LY, Cai CZ, Ji ZL, *et al.* Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucleic Acids Res* 2004;**32**:6437–44.
66. Jayaraj PB, Jain S. Ligand based virtual screening using SVM on GPU. *Comput Biol Chem* 2019;**83**:107143.
67. Spiegel J, Senderowitz H. Evaluation of QSAR equations for virtual screening. *Int J Mol Sci* 2020;**21**:7828.
68. Carracedo-Reboredo P, Linares-Blanco J, Rodriguez-Fernandez N, *et al.* A review on machine learning approaches and trends in drug discovery. *Comput Struct Biotechnol J* 2021;**19**:4538–58.
69. Madhukar NS, Khade PK, Huang L, *et al.* A Bayesian machine learning approach for drug target identification using diverse data types. *Nat Commun* 2019;**10**:5221.
70. Babajide Mustapha I, Saeed F. Bioactive molecule prediction using extreme gradient boosting. *Molecules* 2016;**21**:983.
71. Chen X, Huang L, Xie D, *et al.* EGBMMDA: extreme gradient boosting machine for miRNA-disease association prediction. *Cell Death Dis* 2018;**9**:3.
72. Yu B, Qiu W, Chen C, *et al.* SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting. *Bioinformatics* 2020;**36**: 1074–81.
73. Zhao Z, Yang W, Zhai Y, *et al.* Identify DNA-binding proteins through the extreme gradient boosting algorithm. *Front Genet* 2021;**12**:821996.
74. Ubels J, Schaefers T, Punt C, *et al.* RAINFOREST: a random forest approach to predict treatment benefit in data from (failed) clinical drug trials. *Bioinformatics* 2020;**36**:i601–9.
75. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol Methods* 2009;**14**:323–48.
76. Chen X, Yan CC, Zhang X, *et al.* WBSMDA: within and between score for miRNA-disease association prediction. *Sci Rep* 2016;**6**:21106.
77. Burggraaff L, Oranje P, Gouka R, *et al.* Identification of novel small molecule inhibitors for solute carrier SGLT1 using proteochemometric modeling. *J Chem* 2019;**11**:15.
78. Zhang C, Zhou Y, Gu S, *et al.* In silico prediction of hERG potassium channel blockage by chemical category approaches. *Toxicol Res* 2016;**5**:570–82.
79. Lu L, Qin J, Chen J, *et al.* DDIT: an online predictor for multiple clinical phenotypic drug-disease associations. *Front Pharmacol* 2021;**12**:772026.
80. Luo M, Wang XS, Tropsha A. Comparative analysis of QSAR-based vs. chemical similarity based predictors of GPCRs binding affinity. *Mol Inform* 2016;**35**:36–41.
81. Schurer SC, Muskal SM. Kinome-wide activity modeling from diverse public high-quality data sets. *J Chem Inf Model* 2013;**53**: 27–38.
82. Ma J, Sheridan RP, Liaw A, *et al.* Deep neural nets as a method for quantitative structure-activity relationships. *J Chem Inf Model* 2015;**55**:263–74.
83. Sauvat A, Cerrato G, Humeau J, *et al.* High-throughput label-free detection of DNA-to-RNA transcription inhibition using brightfield microscopy and deep neural networks. *Comput Biol Med* 2021;**133**:104371.
84. Date Y, Kikuchi J. Application of a deep neural network to metabolomics studies and its performance in determining important variables. *Anal Chem* 2018;**90**:1805–10.
85. Arcos-Garcia A, Alvarez-Garcia JA, Soria-Morillo LM. Deep neural network for traffic sign recognition systems: an analysis of spatial transformers and stochastic optimisation methods. *Neural Netw* 2018;**99**:158–65.
86. Eppenhof KAJ, Lafarge MW, Veta M, *et al.* Progressively trained convolutional neural networks for deformable image registration. *IEEE Trans Med Imaging* 2020;**39**:1594–604.
87. Yamanakkanavar N, Lee B. A novel M-SegNet with global attention CNN architecture for automatic segmentation of brain MRI. *Comput Biol Med* 2021;**136**:104761.
88. Abdelrahman L, Al Ghamdi M, Collado-Mesa F, *et al.* Convolutional neural networks for breast cancer detection in mammography: a survey. *Comput Biol Med* 2021;**131**:104248.
89. Jia S, Hu P. ChrNet: a re-trainable chromosome-based 1D convolutional neural network for predicting immune cell types. *Genomics* 2021;**113**:2023–31.
90. Xu Z, Shen D, Nie T, *et al.* A hybrid sampling algorithm combining M-SMOTE and ENN based on random forest for medical imbalanced data. *J Biomed Inform* 2020;**107**:103465.
91. Wishart DS, Feunang YD, Guo AC, *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;**46**:D1074–82.
92. Nardin I, Kollner S. Successful development of oral SEDDS: screening of excipients from the industrial point of view. *Adv Drug Deliv Rev* 2019;**142**:128–40.
93. Rayaprolu BM, Strawser JJ, Anyarambhatla G. Excipients in parenteral formulations: selection considerations and effective utilization with small molecules and biologics. *Drug Dev Ind Pharm* 2018;**44**:1565–71.
94. Rao VA, Kim JJ, Patel DS, *et al.* A comprehensive scientific survey of excipients used in currently marketed, therapeutic biological drug products. *Pharm Res* 2020;**37**:200.
95. Wang X, Li F, Qiu W, *et al.* SYNBIP: synthetic binding proteins for research, diagnosis and therapy. *Nucleic Acids Res* 2022;**50**: D560–70.
96. Kim S, Chen J, Cheng T, *et al.* PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res* 2021; **49**:D1388–95.
97. Lancet T. ICD-11. *Lancet* 2019;**393**:2275.

Downloaded from https://academic.oup.com/bib/article/23/5/bbac160/6582006 by National Science & Technology Library user on 07 October 2022

98. Zahednezhad F, Saadat M, Valizadeh H, *et al.* Liposome and immune system interplay: challenges and potentials. *J Control Release* 2019;**305**:194–209.
99. Davies A, Velickovic P, Buesing L, *et al.* Advancing mathematics by guiding human intuition with AI. *Nature* 2021;**600**:70–4.
100. Gillis J, Pavlidis P. The role of indirect connections in gene networks in predicting function. *Bioinformatics* 2011;**27**:1860–6.
101. Sun M, Zheng Q. Key factors in conformation transformation of an important neuronic protein glucose transport 3 revealed by molecular dynamics simulation. *ACS Chem Neurosci* 2019;**10**:4444–8.
102. Zhang Y, Zhang H, Zheng Q. How chorismatases regulate distinct reaction channels in a single conserved active pocket: mechanistic analysis with QM/MM (ONIOM) investigations. *Chemistry* 2019;**25**:1326–36.
103. Zhang S, Amahong K, Zhang C, *et al.* RNA-RNA interactions between SARS-CoV-2 and host benefit viral development and evolution during COVID-19 infection. *Brief Bioinform* 2022;**23**:bbab397.
104. Zhang S, Amahong K, Sun X, *et al.* The miRNA: a small but powerful RNA for COVID-19. *Brief Bioinform* 2021;**22**:1137–49.
105. Zhang Y, Zheng QC. What are the effects of the serine triad on proton conduction of an influenza B M2 channel? An investigation by molecular dynamics simulations. *Phys Chem Chem Phys* 2019;**21**:8820–6.
106. Zhang Y, Zhang H, Zheng Q. A unique activation-promotion mechanism of the influenza B M2 proton channel uncovered by multiscale simulations. *Phys Chem Chem Phys* 2019;**21**:2984–91.
107. Zhang Y, Zhang H, Zheng Q. What regulates the catalytic activities in AGE catalysis? An answer from quantum mechanics/molecular mechanics simulations. *Phys Chem Chem Phys* 2017;**19**:31731–46.
108. Xue W, Yang F, Wang P, *et al.* What contributes to serotonin-norepinephrine reuptake inhibitors' dual-targeting mechanism? The key role of transmembrane domain 6 in human serotonin and norepinephrine transporters revealed by molecular dynamics simulation. *ACS Chem Neurosci* 2018;**9**:1128–40.
109. Xue W, Fu T, Deng S, *et al.* Molecular mechanism for the allosteric inhibition of the human serotonin transporter by antidepressant escitalopram. *ACS Chem Neurosci* 2022;**13**:340–51.
110. Zhang Y, Ying JB, Hong JJ, *et al.* How does chirality determine the selective inhibition of histone deacetylase 6? A lesson from trichostatin a enantiomers based on molecular dynamics. *ACS Chem Neurosci* 2019;**10**:2467–80.
111. Xue W, Wang P, Tu G, *et al.* Computational identification of the binding mechanism of a triple reuptake inhibitor amitifadine for the treatment of major depressive disorder. *Phys Chem Chem Phys* 2018;**20**:6606–16.
112. Tang J, Fu J, Wang Y, *et al.* Simultaneous improvement in the precision, accuracy, and robustness of label-free proteome quantification by optimizing data manipulation chains. *Mol Cell Proteomics* 2019;**18**:1683–99.
113. Fu J, Tang J, Wang Y, *et al.* Discovery of the consistently well-performed analysis chain for SWATH-MS based pharmacoproteomic quantification. *Front Pharmacol* 2018;**9**:681.
114. Yang Q, Li B, Chen S, *et al.* MMEASE: online meta-analysis of metabolomic data by enhanced metabolite annotation, marker selection and enrichment analysis. *J Proteomics* 2021;**232**:104023.
115. Yang Q, Hong J, Li Y, *et al.* A novel bioinformatics approach to identify the consistently well-performing normalization strategy for current metabolomic studies. *Brief Bioinform* 2020;**21**:2142–52.

Downloaded from https://academic.oup.com/bib/article/23/5/bbac160/6582006 by National Science & Technology Library user on 07 October 2022